403 487

TECHNICAL NOTE NO. 7

under Contract
AF 49(638)-1128

"Inductive Methods in Language
Analysis"

1 May 1963

DDC

MAY 14 1963

A

873 050

THOMPSON RAMO WOOLDRIDGE INC.

RW DIVISION

CANOGA PARK, CALIFORNIA

NO OTS

Best
Available
Copy

⑤ *873 650*

(7) NA
(8) u
(9) NA
(11) 1962
(12) 25 p.
(13) NA (14) TN 7
(16-19) NA
(20) u
(21) NA

⑥ PROBLEMS OF METHOD
IN THE
COMPUTER PROCESSING OF CULTURAL DATA[1]

⑩ Paul L. Garvin
Thompson Ramo Wooldridge Inc.
RW Division

The purpose of this paper is to consider some crucial aspects
of the problems connected with the processing of cultural data on
computing equipment.[2]   The point of view taken here is that in plan-
ning the application of this equipment to a new field of study, certain
questions are more closely related to the characteristics of com-
puter applications than to the nature of the problems that are to be
studied.

One set of questions has to do with the use of the computer
as a tool for information processing, and more particularly with the
input and output problems that it raises: (1) what type of information
is required in order to make the use of the equipment possible, that
is, what is needed to make a problem computable? (2) what type of
information can be expected as the result of computer use?

1

The second type of question has to do with the way in which the equipment is used, that is with the actual computer operations: what is the relation of the characteristics of computer programs to the characteristics of the data?

In view of the very limited experience which exists to date in the use of computers for the study of culture, these questions can at present only be considered in the abstract and in the most general terms. At best we can hope that by a consideration of these questions in the light of the experience of a related field such as linguistics we may gain some further insights. Thus, while we are able to examine a variety of possibilities and formulate some relevant questions, the answers will not be available until after the problems have been worked out. Therefore we will make suggestions in terms of our conception of what can be expected, and state what we consider relevant for future research. In doing this, we shall try to maintain a clear distinction between what we consider only slightly hypothetical, and what we consider entirely speculative.

We shall begin by answering the questions raised by the information-processing aspects of computer use, first in general terms, then in the light of the experience of linguistics. We shall conclude by some observations on the characteristics of computer programs and how these may possibly relate to the study of culture.

* * *

The computer has often been called a tool for the manipulation of symbols with high speed of operation and extensive memory capacity. Its proper use can, as is well known, effect large savings of time and money. But it is only when significantly more is gained by the use of computers than is required for making the problem computable that the use of the equipment is justified.

2

The first question which we asked in the introduction—namely, what type of information is needed to make a problem computable— has to do with the proper use of the computer. The second question— namely, what type of information can be expected—becomes more interesting if it is asked in a more specific form: what does the use of the computer contribute beyond the possible advantages of speed and economy? This is particularly significant in research applications, where savings of time and money are not a primary intellectual aim.

Two requirements have to be met before a computer can be used: (1) the data have to be formatted before the computer will accept them as input, that is, they have to be formulated in appropriate discrete symbols and presented in a suitable order; (2) a program must first be outlined (most often in the form of flowcharts) and then written out in some computer code.

An extensive preparation is thus a prerequisite for computer processing. The significance of this preparation is that it requires an uncompromisingly explicit and rigorous formulation of the under- lying assumptions as well as of the aims of the study, since in order to minimize costly errors during the actual processing, the implica- tions of each research decision must be traced out beforehand in as much detail as possible. This forces hidden assumptions to become overt and accessible to evaluation, which in turn may force their re- vision or abandonment. It requires that aims be stated specifically enough to be attainable in fact as well as desirable in theory, and hence may force their reformulation.

I should like to illustrate these points by my recent work on the design of a computer program for automatic linguistic analysis. [3] I defined automatic linguistic analysis as the processing of a suitably large body of text in a given language by a computer program which would not be based on information particular to that language, but only

[3] Paul L. Garvin, "Automatic Linguistic Analysis—a Heuristic Problem," Proc. Internat. Conference on Machine Translation and Applied Language Analysis, Teddington, England, 1961, in press with H. M. S. O.

on general linguistic assumptions. The purpose of this program would be to produce as output a description of the language represented by the text.

As I examined the routines which I had devised for the automatic linguistic analysis program, it became apparent that some of them were not based solely on general linguistic assumptions. They presupposed for their operation the presence, in the text, of certain structural characteristics (such as the relative significance of word order) which, while not limited to a single language, are not universal and therefore are not included in the general linguistic assumptions to which the program was supposed to be limited. Thus, these routines reflected a hidden assumption which, when made explicit, was shown to be typological. The program, instead of being suited for processing a given language (in the sense of any given language), had turned out to be restricted to the processing of only a language of a given type (namely, one—such as Chinese or English—in which word order is significant). The typological assumption was added to the specifications of the program which accordingly were revised to include the restriction to languages of a particular type. Rather than invalidating the research, uncovering a hidden assumption thus led to a new insight: an increased awareness of the significance of word order as a typological criterion, and of the relevance of language typology to linguistic analysis.

Similarly, the aim of the program was initially set forth simply as one of automatic linguistic analysis. As the program was considered in more detail, it became necessary to take into account the significant differences between the two major aims of linguistic analysis: segmentation into discrete units, and analysis of the distribution of these units:

"Linguistic segmentation is the first step in the analysis of raw text—that is, of spoken messages recorded from native informants. Segmentation procedures are based on the relation between the form (i. e. , the phonetic shape) and the meaning (in operational terms, the translation or possible paraphrase) of the message. Their

4

mechanization thus would require the comparative processing of two inputs — one representing the phonetic shape of the raw text, one its translation or paraphrase.

A program designed for a single rather than a dual input hence cannot be expected to accomplish segmentation. On the contrary, it requires an input which has in some prior way been segmented into elements of a unified functional type. Given such a previously segmented input, the program can, however, be expected to accomplish a distributional analysis of the elements that are found to recur in the input text. "[4]

Consequently, the aim of the program was reformulated more specifically as that of distributional analysis, leading to a distributional description.

We have already noted that the preparation for computer processing is not only a prerequisite for computer use, but has an important effect on the research design. Some or all of the questions to which the research is addressed may be answered by the preparation, or part of it, before any actual computer runs have taken place. In the case of my program for automatic linguistic analysis, which so far only exists in the form of a research paper outlining it, one important question has already been answered just by giving a verbal description in sufficient detail. It is a question of considerable significance to the theoretical foundations of linguistic research, namely, whether the automation of descriptive linguistic procedures is conceivable.

Let me now discuss the function of formatting and programming in the research design somewhat more specifically.

Formatting, as we said above, consists in the formulation of the input data in appropriate discrete symbols and in their presentation in a suitable order.

---

[4] Op. cit. in fn. 3.

5

From the standpoint of formatting, continuous written text in a natural language (such as English) differs from other types of anthropological input data in that it comes already formatted. It consists of appropriate discrete symbols—letters, punctuation marks, spaces. The symbols are arranged in a suitable order—that determined by the rules of the language. Only some editorial conventions are usually required to insure proper keypunching. [5]

Data other than text, on the other hand, have to be formatted as part of the general research design. The requirement of discrete symbols forces the determination, or postulation, of discrete input units. The requirement of ordering forces the determination, or postulation, of at least some relations (they may be as elementary as the temporal succession implicit in the consecutive inputting of the data). This aspect of computer use makes the consideration of the segmentation of cultural data into units mandatory; it shifts the emphasis from the search for connections to the search for units.

Madeleine Mathiot has in a recent paper stressed the importance of two categories of units relevant in the description of behavior, behavioral units and analytic units. She specifies them as follows: "... we consider that behavioral units are perceptually defined segments of behavior, i.e., they are derived from behavior by observation. We consider that analytic units are conceptually defined segments, i.e., they are inferred by operational rules from some previously given entities." [6]

A nonlinguistic example of a behavioral unit in the above sense would be an observed ceremony, an example of an analytic unit, a role.

----

[5] As is well known, keypunching (or some other comparable manual operation) is at present the only way of mechanically adapting the data for input into a computer.

[6] Madeleine Mathiot, "The Place of the Dictionary in Linguistic Description: Problems and Implications," in press with Language.

In terms of the above, the first question that arises in connection with formatting is: is some suitable behavioral or analytic unit available in the problem area under consideration, or is the question of units faced for the first time in connection with the need for the input units required by the format.

Thus, using the above example of a behavioral and an analytic unit, for a particular problem it may be reasonable to represent, say, each observed ceremony, or each role, by a given discrete symbol and input these symbols in an order deemed appropriate for purposes of the problem. It can be imagined that for a given problem area more than one type of unit might be available, in which case the further question would arise as to which type of unit is most appropriate to the problem, and conceivably even which of several equally appropriate types is best suited for formatting. With non-textual data, however, it is much more likely that the opposite will be the case — namely, that no previously established analytic or behavioral unit is given for the problem area under study. It will then be necessary to stipulate appropriate input units without the reliance on a behavioral or analytic unit. It is this latter case which makes formatting of greatest interest to the study of the segmentation of cultural data. Formatting then becomes a primary means of segmentation and input units become analytic devices in their own right, independently of behavioral units and analytic units.

The important thing to note here is that the only requirement imposed by formatting is that the input units be discrete — in all other respects, they may be arbitrarily chosen, as long as the format fits the problem. This seemingly allows the determination of input units at the analyst's convenience, but in actuality forces the consideration of the aim of the research and the formulation of some segmentation criterion in terms of this aim. At the same time, the subsequent computer processing can be expected to provide a means of verification: will the results obtained by inputting the units chosen on the basis of a given criterion be satisfactory in terms of our expectations? Thus, if we assume, for purposes of a given problem, that X, Y, Z,

7

are separate and individual traits, in that order, and input them accordingly, will the program help us solve our problem, or will we have to revise our input assumptions before the program will yield results that we can use?

Formatting thus requires and—once accomplished—provides a means of, segmenting and ordering the data. The program for which the format is designed plays a similar part with respect to the problem: it requires and—once accomplished—provides a means for, structuring the problem. Because of the explicitness required by the use of the computer, the problem must be decomposed into small components. This is facilitated by a symbolization which is not only suited to, but requires, such a breakdown, namely, the graphic display of the step-by-step logic of the problem in the form of a flowchart,[7] a display which is so detailed that it does not tolerate skipping any steps.

From the standpoint of possible applications in the study of culture, we can distinguish two broad categories of computer programs: statistical and nonstatistical.

Statistical programs are at present applied mainly to the type of problems encountered in historical anthropology or historical linguistics. The use of computers is justified if the bulk of the data permits full utilization of the speed and capacity of the equipment. It differs from the use of statistics without computers by a more rigorous formatting requirement.

As has been apparent from the discussion so far, the emphasis in the present paper is on nonstatistical applications. In these applications, the role of the computer has become that of a logical tool. The significant property of the equipment no longer is its well-known

---

[7] A detailed verbal description may sometimes take the place of a flow diagram, as in the case of our own automatic linguistic analysis mentioned further above. Nevertheless, we consider flowcharting, rather than actual programming, the skill most useful to the prospective new computer user.

capacity for processing large bodies of data very rapidly. On the
contrary, the emphasis is on the property of computer programs
of carrying out a given set of instructions, and only those instruc-
tions, to the letter. This means, in effect, that a computer run
will show exactly what are the consequences of a particular set of
assumptions, without the intuitive improvements that the human
investigator is tempted to introduce in following up his premises.
Thus, whatever hidden assumptions and inconsistencies may have
escaped detection during the preparation of the run, will now be
revealed.

To illustrate a nonstatistical computer application, we pre-
sent the flow diagram required if we wanted to design a computer
program for applying the definition of a social science concept to
particular cases.

The definition is that of the organization of a biological or
social system by Ruesch and Bateson:

"The concept of organization ... in spite of much discussion
remains imperfectly defined, but as the word is usually understood
it seems to have the following components: a system is regarded
as organized (a) if it consists of several active entities — cells,
organs, individuals, and so forth; (b) if among these entities certain
similarities and differences occur; (c) if communication occurs be-
tween the active entities; and (d) if the continuation of the system
in a steady state depends upon this communication and upon these
similarities and differences of activity."[8]

The purpose of a program based on this definition might be to
answer, for any given system, whether or not it is organized. The
flowchart for this conceivable program is shown in Figures 1
and 2.

---

[8] Jurgen Ruesch and Gregory Bateson, "Structure and Process in
Social Relations," Psychiatry 12. 112 (1949).

ENTER WITH SYSTEM

1 DOES THE SYSTEM CONSIST OF SEVERAL ENTITIES? — YES

2 ARE THE ENTITIES ACTIVE? — YES

3 ARE THE ENTITIES CELLS? — YES

NO

5 THE SYSTEM IS NOT ORGANIZED

4 ARE THE ENTITIES ORGANS? — YES

NO

6 ARE THE ENTITIES INDIVIDUALS? — YES

NO

7 ARE ALL THE ENTITIES IDENTICAL WITH EACH OTHER? — YES → 8 UNDEFINED CONSEQUENCE

NO

9 ARE SOME OF THE ENTITIES IDENTICAL WITH EACH OTHER? — YES → 10 UNDEFINED CONSEQUENCE

NO

11 ARE THERE ANY SIMILARITIES BETWEEN THESE ENTITIES? — YES → 12 DEFINE TYPE OF SIMILARITY

NO

13 ARE THESE SIMILARITIES OF THE DEFINED TYPE? — YES

NO

14 DEFINE TYPE OF DIFFERENCE

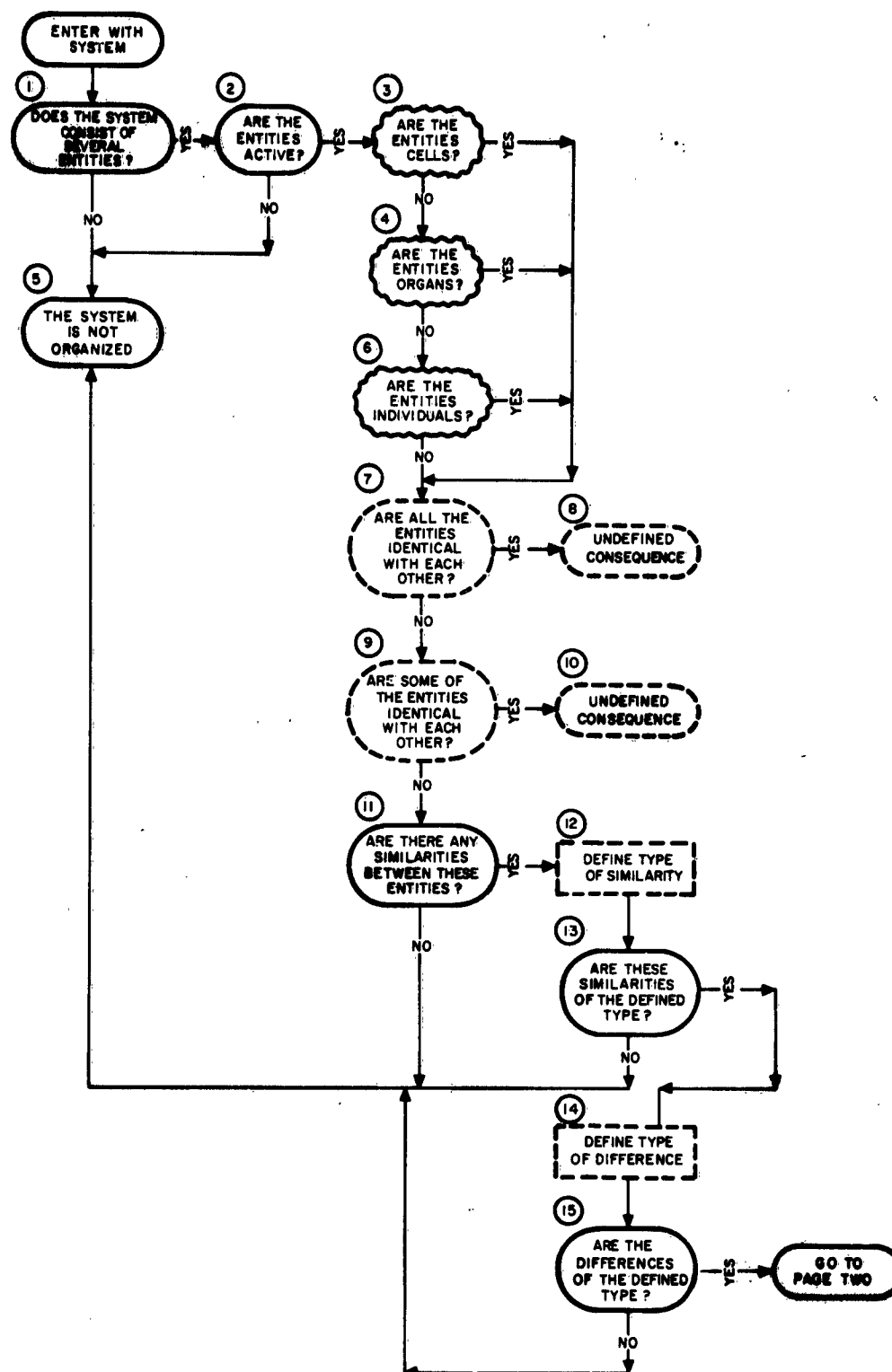15 ARE THE DIFFERENCES OF THE DEFINED TYPE? — YES → GO TO PAGE TWO

NO

Figure 1. Flowchart of System Organization
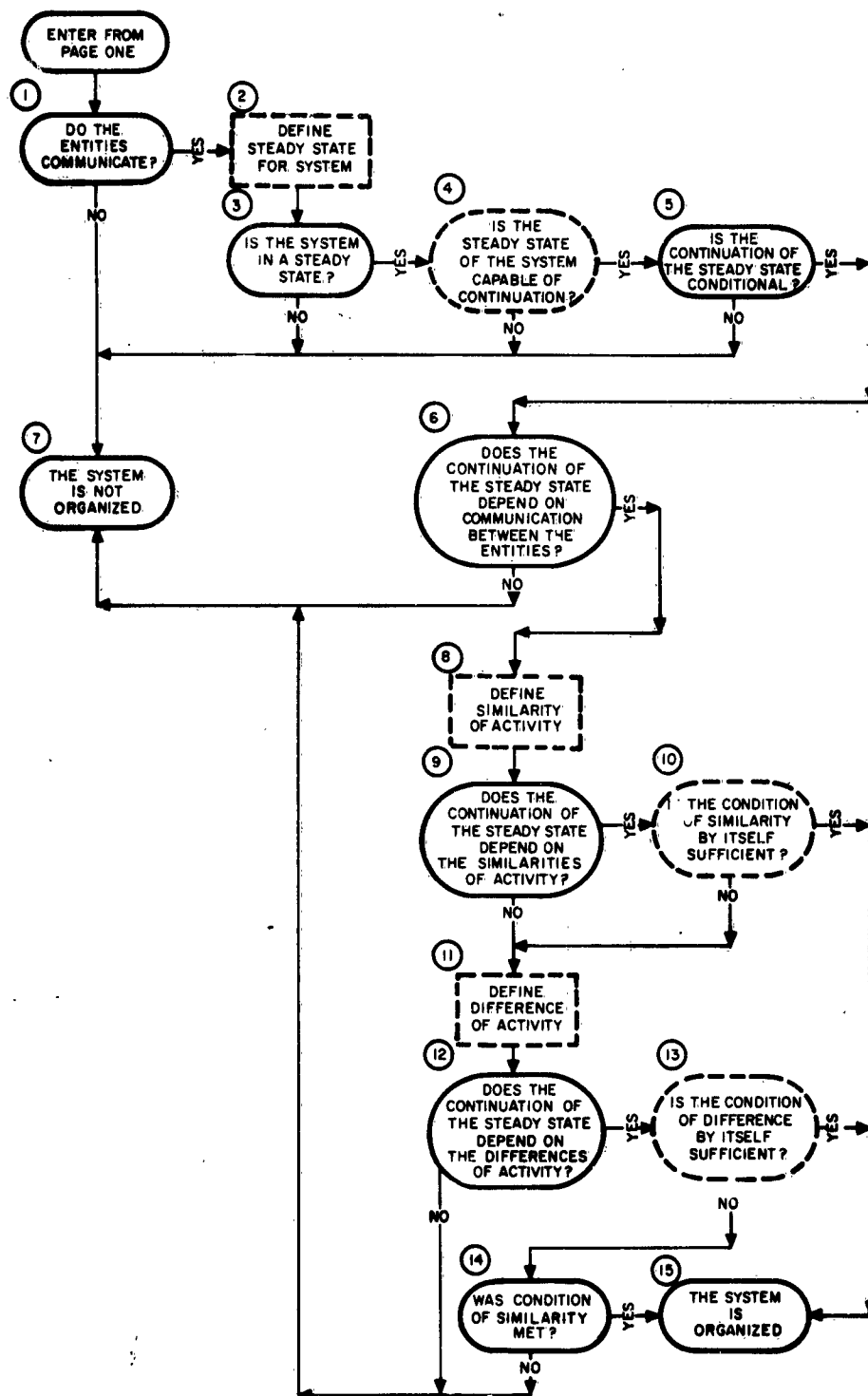(Based on Ruesch and Bateson) (Page 1)

10

Figure 2. Flowchart of System Organization
(Based on Ruesch and Bateson) (Page 2)

11

Before discussing this flow diagram, a few remarks on flow-
charting conventions are indicated. As can be seen from the figures,
a flowchart consists of a series of "boxes" connected by arrows.
The former contain the explanations, the latter indicate the direction
of the "flow." Two general conventions are worth mentioning:[9]
(1) boxes with rounded edges enclose entrances, exits, and questions,
boxes with pointed edges enclose instructions or information; (2) "yes"
arrows start out horizontally to the right, "no" arrows start out
vertically downward. We introduced a few special conventions for
our particular flowchart: we have numbered all the boxes consecutively
on both pages, and we have indicated the relation between the flowchart
and the original definition by drawing different borders around the
boxes. Solid borders indicate questions and information taken directly
from the verbal definition; broken lines bordering a box indicate
questions or information not explicitly contained in the text but im-
plied by it as hidden assumptions necessary to the logical flow; wavy
lines bordering a box indicate questions or information contained in
the text and deemed redundant.

The questions on the flowchart are addressed to the input, which
means that the input data must be chosen and formatted so as to pro-
vide the information needed to answer the questions. In our case, the
notation "Enter with system" in the unnumbered box at the top of
page 1 indicates that the input to the program represented by our flow-
chart would be a system. This means that the system which we would
wish to study by means of our program would have to be specified
precisely enough so it both could be formatted for input, and would
provide the information required to answer the questions on the flow-
chart.

The flowchart shown in Figures 1 and 2 represents one inter-
pretation of the original verbal definition. We can now compare our

[9]Cf. "Proposed Standard Flow Chart Symbols," Communications of
the Association for Computing Machinery, vol. 2, no. 10, pp. 17-8
(October 1959).

flowchart to this interpretation. There are two aspects in which
we consider such a comparison meaningful: (1) we can compare
the overall structure of the flowchart to that of the verbal definition;
(2) we can compare certain details of the flowchart with the corre-
sponding portions of the verbal statement. It is important to note
that a different interpretation of the original definition will result
in significant differences in the flowcharting. The differences in
flowcharting due to a different interpretation of one portion of the
definition are shown in Figure 3.

    First, a brief comparison of the two structures. The verbal
definition contains four parts, labeled (a), (b), (c), (d) respectively
by the authors. These parts correspond to boxes on the flowchart
as follows:

        Part (a) - boxes 1, 2, 3, 4, 6, on page 1
        Part (b) - boxes 7, 8, 9, 10, 11, 12, 13, 14, 15, on page 1
        Part (c) - box 1 on page 2
        Part (d) - boxes 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14,
                on page 2

    Note that the different parts of the verbal definition correspond
to rather unequal portions of the flowchart: part (c) corresponds to
a single box, while part (d) corresponds to as many as 12 boxes. We
might want to conclude from this that the logical structure of the
original definition is uneven, in the sense that the conditions set forth
in the different parts of the definition are not of the same order.
This interesting characteristic of the verbal statement both permits
and encourages us to explore whether the unevenness in structure
constitutes a logical flaw in the definition or reflects a significant
and hitherto unnoticed property of the defined object. In the details,
the flowchart differs from the verbal definition primarily by being
more explicit, as is shown by the special flowcharting conventions
that we have adopted. By way of illustration we want to point out the
hidden assumptions implicit in part (b) of the definition which reads:
"if among these entities certain similarities and differences occur."

13
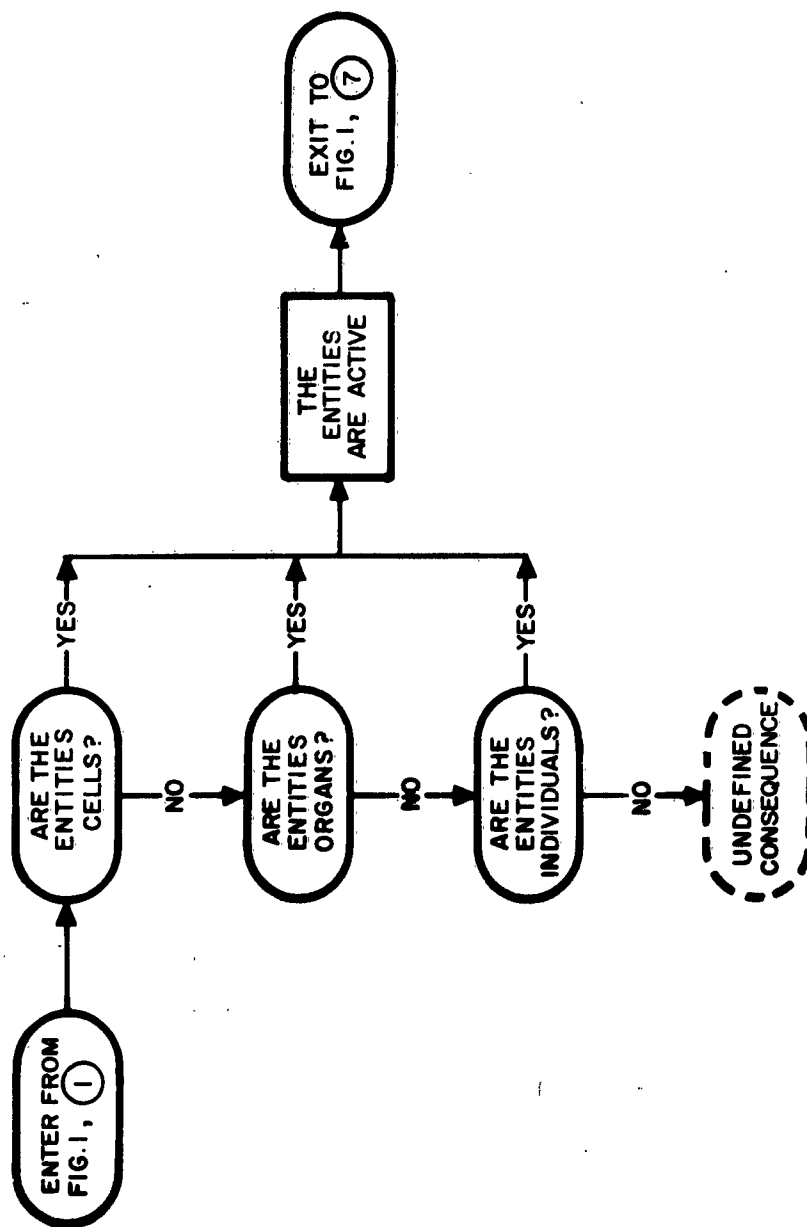
Figure 3. Alternate Branching of Flowchart of System Organization (Based on Ruesch and Bateson)

Boxes 7 through 10 of page 1 of the flowchart indicate that this statement either implies the assumption that neither all nor some of the elements are identical, or ignores the possibility of a condition of identity altogether (this is indicated by the notation "Undefined consequence" in boxes 8 and 10). Boxes 12 and 14 indicate that the term "certain" in the verbal statement implies a particular type of similarity or difference which requires specification.

Finally, we want to point out that boxes 2 through 6 on page 1 of the flowchart are based on interpreting part (a) of the verbal definition to mean that "cells, organs, individuals, and so forth" are merely redundant elaborations of the requirement for "active entities." An alternative interpretation is shown in Figure 3 which is based on assuming that "cells, organs, individuals, and so forth" are meant to be an incomplete definition of the requirement for "active entities." Note that while this interpretation removes the redundancy, it creates an "Undefined consequence" due to the statement "and so forth."

* * *

The basic distinction between textual and nontextual data, which we mentioned in the discussion of formatting further above, applies to the entire field of the computer processing of cultural data.

Although we wish to concentrate upon the problems of processing nontextual data, we shall first give some attention to the question of textual data.

The textual data of the anthropologist c n be divided into two major categories: (1) text produced by the cultu.e under study, such as folklore; (2) text produced by anthropologists, such as ethnographic descriptions.

The processing of text produced by the culture raises the question of computer applications for purposes other than simple tabulation, which, while it may unquestionably be useful, does not

require further discussion. In other words, can the computer be used for purposes of analysis, and more specifically, for an analysis that is primarily cultural rather than linguistic? At our present state of knowledge we are not yet in a position to set forth the conditions for a cultural analysis of text for more than the automatic production of concordances.

In corcordance automation, text size and cost are significant considerations. Unless a simple general concordance program is available ready for use in some program library, the amount of text at hand for a particular culture may not warrant the time or effort required to write such a program. Once a concordance has been produced, it is a useful added tool for analysis — it assembles the pieces of the text in a form convenient for inspection. But it must be remembered that a concordance doesn't of itself constitute an analytic result.

Text produced by anthropologists, that is, field notes or other ethnographic documents, are from the standpoint of computer processing merely technical documents in the particular field of anthropology. Hence, the question of treating them automatically is simply a question of information processing — an automated information retrieval activity or an automatic abstracting activity, dealing with cultural anthropological documents rather than with the more usual documents of physics or chemistry. The anthropologist here would become the customer of an automatic system the same as the physical scientist or librarian or government official. The serious question which arises here is whether or not the enormous cost of such automatic systems, particularly in view of their present imperfections, is warranted by the comparatively limited needs of the profession.

We take a more positive view of the opportunities for research on the structure of cultural anthropological terminology afforded by the requirements of information retrieval or automatic abstracting. This promises to be productive and to lead to interesting insights, but we consider the study of terminology a linguistic rather than a cultural problem.

* * *

16

Turning now to the processing of nontextual data, we first want to repeat that we are interested in nonstatistical processing. We also wish to stress again that under nontextual data we include all data that do not consist of connected text. This means that data which consist of isolated verbal material without the connectedness provided by linguistic relations are classed as nontextual, since our differentiative criterion is the ordering imparted to the data by the structure of a natural language, rather than the verbal substance as such. It is also worth noting that our distinction of textual and nontextual does not coincide with the usual differentiation of verbal and nonverbal behavior: data on either form of behavior can be textual or nontextual.

The nonstatistical computer processing of nontextual data in an anthropological frame of reference has, to our knowledge, so far not been attempted on a serious scale. We are therefore justified in considering a related frame of reference, that of linguistics, and examining its possible bearing on the question at hand. In this consideration, we shall take as our point of departure an earlier paper in which we suggested that there might be three basic degrees of computer participation in linguistic research.[10] We proposed that the lowest degree is linguistic data collection, an intermediate degree is the verification by a computer program of the results of linguistic research obtained through other means, and the highest degree is testing the validity of linguistic method by a computer program. As an important example of the first degree, we used the automatic compilation of a concordance. As an example of the second degree we used machine translation, and as an example of the third degree we used automatic linguistic analysis, in particular, our own conception of it which is one of distributional analysis-(first cited further above to illustrate the effects of computer processing on the assumptions and aims of research, see pp. 3-5 and footnote 3).

---

[10] Paul L. Garvin, "Computer Participation in Linguistic Research," Language 38. 385-9 (1962). (Originally presented at the Wenner-Gren Symposium).

The reasons for considering that these are different degrees of computer participation are as follows: In the compilation of a concordance or other means of language data collection, the computer is used in what is essentially a bookkeeping and filing function. While it is true that all computer operations can ultimately be reduced to a form of bookkeeping and filing, it is equally true that many computer programs have a logical structure far transcending that of the bookkeeping and filing components of which they are made up. This certainly applies to the computer programs used in machine translation and automatic linguistic analysis. A machine translation program in the context of linguistic research serves as a tool to verify the correctness of a particular analysis: if the analysis on which the translation program is based is correct, the program will produce acceptable translation, and if not, it will not. Finally, it is clear that in the case of automatic linguistic analysis the program will carry out, with the necessary logical consistency, the analytic instructions built into it. Clearly, if these are good instructions the output of the program will be acceptable, and if they are not, it will not.

We can now ask whether the linguistic frame of reference provided by these degrees of computer participation can help us in developing a systematic approach to the nonstatistical processing of nontextual data. As a first approximation, we can explore whether the three processes which we cited as examples of the three degrees — namely, concordance-making, translation, and distributional analysis — are in any way applicable to nontextual as well as textual data. The concrete question will be: is it possible to conceive of a concordance, a translation, a distributional analysis, of nontextual data. The consideration of this question will of necessity have to be even more speculative than the discussion so far. But even if an answer turns out to be impossible to give or trivial, the question is worth asking because it may lead us to an understanding of the nature of cultural data in a new light.

From this standpoint an interesting division suggests itself. Upon brief consideration, it will be apparent that of all the varieties

18

of nontextual data, music may lend itself most readily to the above named processes. This is not surprising, since music is comparable to language in the sense that it, too, forms a separate and self-contained system; furthermore, music permits the use of a discrete notation which can readily, though not at this stage mechanically, be transformed into computer input.

It is quite conceivable that a musical concordance can be constructed. The requirements for such a concordance would be: (1) some clearly delimitable units, the environments of which are to be set forth in the concordance; (2) some frame unit, if the environments are to be defined by more than just a certain mechanically determined number of units adjacent to each side. The required units exist in music in the form of its well-known rythmical stretches. As in the case of language, we would not consider a musical concordance an analytic result, but rather a tool for further analysis. It would ex.eed the scope of this paper to speculate how such a tool might be used in ethnomusicology.

Translation can likewise be envisioned in the case of music. It is common knowledge that musical pieces can be transposed from one scale into another (for instance, from diatonic into pentatonic or conversely), or that scores can be rewritten for different instrumentation (for instance, from chamber quartet to symphony orchestra). From an anthropological standpoint, it would be important to differentiate between cross-cultural and intracultural forms of translation. It would also be interesting to inquire what particular property (such as, for instance, the same melody) forms the basis on which two pieces of music would be considered translations of each other by a given listening or performing community. We can conceive of the meaningful application of such an approach to both descriptive and comparative problems.

Finally, we can envision a form of musical analysis similar to distributional analysis in linguistics, namely, the analysis of the occurence pattern of smaller musical units, such as bars, within larger musical units. This might be an interesting question to

19

consider in its own right, and the methods and results of distributional analysis might be compared to those of other forms of analysis.

We are convinced that any one of these operations could be automated. The significant question, in our opinion, would be not only how automation could best be implemented, but what research advantages would be gained from automation that go beyond the results which could be obtained by attempting concordance-making, translation, and distributional analysis of music "manually."

In the case of a musical concordance (just as in the case of language concordances) one could state that automation only makes sense for a body of musical data large enough to warrant the use of computing equipment.

The interest of automatic translation in music would be similar to that of automatic language translation. A program which would automatically transpose one form of music into another would produce the logically consistent results of certain assumed rules of transposition, which would have to be stated with the requisite explicitness. The output of the program would provide a mode of verification comparable to that obtained in language translation: by performing this output, the logical consequences of the transposition rules could be listened to and the adequacy of the rules could be judged.

We can most readily conceive of automatic musical analysis along lines similar to automatic linguistic analysis, that is, as a computer program for distributional analysis (cf. pp. 3-5 and op. cit. in fn. 3). It may be possible to consider the automation of the better known forms of musical analysis; we are not in a position to judge whether the procedures of musical analysis have ever been set forth in sufficient detail to permit computer use. Here again, the most conspicuous advantages would be explicitness and logical consistency.

We may now go on to speculate about the possible application of concordance-making, translation, and distributional analysis to nontextual data other than music. Such nontextual data might consist

20

of symbolic notation such as that of kinesics[11] or choreography, or of direct representations such as motion picture film. While the notations are discrete to begin with, direct representation raises the problem of segmentation into the units required for the desired applications. It seems fairly obvious in the case of film, for instance, that such technologically given segments as individual frames do not have cultural significance.

The problem here is a direct correlate of the more general problem of the segmentation of the nonverbal behavior which is represented by the nontextual data. Once we are able to provide the necessary segmentation, we may look upon this nonverbal behavior as a temporal sequence of segments similar to the linguistic or musical units in their temporal sequence. We can then consider the application of concordance-making, translation, and distributional analysis to this sequence of behavioral segments.

As in the case of musical data, it is clear that a concordance of nonmusical nontextual data is conceivable, given the appropriate segmentation of behavior. As before, it is worth asking question of the use to which such a concordance could be put in analyzing cultural behavior. Whether properly segmented data will be available in sufficient quantity to warrant automating a concordance can at present not even be asked.

In considering the translation of nonmusical nontextual data several interesting questions suggest themselves. Thus, for instance, the question of the comparison constant for determining what segment of behavior in one culture is the translation of some equivalent behavior in another culture has to our knowledge as yet not even been posed. As a very crude approximation one might assume for instance, that the famout list of needs given by Malinowski is a set of such comparison constants,[12] and that behavior in one culture in response to a given such need under a certain statable set of physical

---

[11] Ray L. Birdwhistell, Introduction to Kinesics, Louisville, 1954.

[12] I am indebted to M. G. Smith for this suggestion.

conditions, is the translation of behavior in another culture filling that same need under comparable physical conditions. For instance, behavior at a particular meal in one culture may be considered the "translational equivalent" of behavior at a comparable meal in another culture. It might be instructive to view certain aspects of accultura-tion as instances of "cultural translation."

In this light, it might not be too far-fetched to speculate about a "cultural translation" program which might have the aim of verifying a hypothesis about culture change. The program could be based on the rules of culture change as set forth by the hypothesis; it could be made to operate on a set of data drawn from a situation the outcome of which is historically known. The program could be expected to -devise the logical consequences of the effect of the rules of change on the original situation. A comparison of the output of the program with the historically known outcome might contribute to confirming or infirming the hypothesis. Conceivably, the same program could be applied to more than one situation, serving to test a more ambitious hypothesis. Needless to say that the many procedural safeguards required to insure the validity of such an approach can not even be foreseen at present.

In regard to a distributional analysis of cultural behavior out-side of language and music, the basic questions are again worth posing: that of the units to be examined, and that of their distributional frame. It is conceivable, for instance, to consider the particular separable portions of a ceremony the units under consideration, and consider the ceremony as a whole, the distribution frame for these units.

Something along these lines has been suggested by Pike in his description of the football game and the breakfast in his theoretical discussion of culture.[13] The rather cool reception that Pike's approach has found among culturally oriented workers[14] shows, if

_____

[13] Kenneth L. Pike, Language in Relation to a Unified Theory of the Structure of Human Behavior, part 1, pp. 44-63 (Glendale, 1954).

[14] Cf. Stanley Newman, review of op. cit. in fn. 13, IJAL 22.84-8 (1956), particularly p. 87: "At the present time, however, the formulation of a theory to coordinate the behavioral sciences is scarcely a task which a single individual could be expected to perform successfully."

22

nothing else, the difficulty of applying the analytic concepts of lin-
guistics to nonverbal culture. Nonetheless, the question of distri-
butional analysis seems worth posing,[15] although at this stage, it
seems premature to speculate about its automation.

* * *

We are now ready for our most ambitious speculation: we
propose to consider certain problems of the theory of culture in the
light of a basic characteristic of computer programs. We are think-
ing of the fundamental debate in cultural anthropology regarding the
concepts of function, structure and process. We will, however,
limit ourselves to the latter two concepts, since we are not as yet
able to suggest an interesting way of speculating about the concept
of function in the frame of reference which we are proposing. The
need for a further clarification of the concepts of structure and proc-
ess becomes evident once one leaves such well traveled areas as kin-
ship in the case of structure, or ethnohistorical change in the case
of process.

We base our speculation on the observation of some gross
similarities between the distinction of structure and process on the
one hand, and the distinction of two basic programming techniques,
namely, table lookup and algorithm, on the other.

We can illustrate the difference between table lookup and
algorithm by an elementary example. Given the task of multiplying
two by four, we can either add two four times to itself—which is the
algorithmic approach, or look up 2 × 4 in a multiplication table.
Even in this elementary context, algorithm and table lookup are not

---

[15] Stanley Newman, ibid.: "... for example, his concept of 'spots and
classes,' an elaboration of the substitution-frame procedure, appears
sufficiently practicable and potentially useful to merit testing in the
nonverbal area of behavior."

mutually exclusive but complementary: given the problem of multiplying numbers greater than those contained in the multiplication table, we use an algorithm to decompose the problem into a series of elementary multiplications, the results of which can be looked up in the multiplication table (although by now we have memorized the table and look it up in our memory), and we combine these part results by an additional algorithm to obtain the final result.

A table in a present-day computer program can become quite complex, to the extent of allowing the analogy with a structure; so can an algorithm, to the extent of permitting the analogy with a process. Although the difference between table lookup and algorithm is somewhat more clearly defined than that between structure and process in anthropology, we were able to point ou on the elementary level of our initial example that even here the two opposites are not always clearly separated, although they are more precisely definable. In any reasonably complex computer program, there will be algorithms containing tables or calling for tables, and a large-scale table-lookup scheme will ultimately require an algorithm for finding one's way around in a table.

At one time in linguistic computer applications there was a serious dispute about the advantages of a predominantly algorithmic versus a predominantly table-lookup approach. At present more general considerations of efficiency are applied, and algorithm and table lookup are no longer looked upon as mutually exclusive alternatives. We may want to carry this conception over into our analogy with cultural anthropology and consider that structure and process likewise are not necessarily mutually exclusive alternatives but can be considered as two aspects of the same phenomenon.

We now come to the high point of our speculation. Let us take our analogy seriously to the extent of adopting the oversimplifying assumption that structure is table lookup and algorithm is process, and that table lookup and algorithm are always clearly distinct. We are now in a position to use table lookup and algorithm as operational definitions of structure and process and consider that anything in a

24

cultural description that lends itself to a table-lookup approach is structure, anything that lends itself to an algorithmic approach is process, and finally anything that lends itself to both approaches contains elements of both structure and process.

The final speculation is not completely unrealistic. It is possible to conceive of a computer program simulating, for instance, culture conflict along the lines of the computer programs that are now being written for management games (i. e., the simulation of business competition for purposes of predicting the outcome of management decisions). It would then be possible to consider that, in this conflict simulation, the elements which enter into a stored table represent the structures of the conflicting cultures, and the rules of which the algorithm consists represent the relevant processes. Similarly to what was suggested in the acculturation example further above, the outcome of the simulation program could be compared with an observed outcome of a conflict, and thereby serve to verify the assumptions built into the program. But the aspect which we want to stress in the present context is the operational consideration of the question of structure and process allowed by our technology in the case of the example of culture conflict. Our decision to store the structures in tables and to reserve the algorithm for the actual process of conflict may give us a means of isolating process from structure and of studying the two separately. By varying the content of our tables and our algorithm we may perhaps be able to achieve a controlled variation of both structure and process. By comparing the outputs of these variant programs to each other and to the observed outcome used for purposes of verification, we may conceivably increase our understanding of the elements of structure and process represented by the variations which we introduced into the tables and algorithms.

Needless to say that it is premature to speculate about how such a series of computer programs could be implem which would be both realistic from the stand culture, and manageable in the light of wh and programming.